行政院國家科學委員會專題研究計畫成果報告

Semiparametric Analysis of Two-Sample Censored Data

計畫編號:NSC 89 - 2118- M- 032- 021

執行期限:89年8月1日至90年7月31日

主持人: 林千代　　執行機構及單位名稱: 淡江大學數學系

計畫參與人員: 吳正新及凌雨瑩　　執行機構及單位名稱: 淡江大學數學系

中文摘要

我們提出一個演算法來取得二組樣本受限資料之半參數概似函數的極大值.

關鍵詞: 半參數概似函數.

Abstract

An iterative algorithm is proposed to maximize the semiparametric likelihood on the two-sample censored samples.

Keywords: semiparametric likelihood function

1 Summary

Suppose X^0 and Y^0 represent the lifetime random variables of two samples from two distributions F_1 and F_2 respectively, and t_j represents the distinct observed values among $n_1 + n_2$ observations. Let

$$r_{1j} = \sum_{i=1}^{n_1} I(X_i^0 = t_j),$$

$$r_{2j} = \sum_{i=1}^{n_2} I(Y_i^0 = t_j),$$

and $r_j = r_{1j} + r_{2j}$ for $j = 1, \ldots, h$. Then the likelihood function based on the complete data is

$$\begin{aligned} & L(\theta, \boldsymbol{p}) \\ = & \prod_{i=1}^{n_1} dF_1(x_i^0) \prod_{i=1}^{n_2} \frac{w(y_i^0, \theta) dF_1(y_i^0)}{\int_0^\infty w(y, \theta) dF_1(y)} \\ = & \prod_{j=1}^{h} [dF_1(t_j)]^{r_{1j}} \prod_{j=1}^{h} \left[\frac{w(t_j, \theta) dF_1(t_j)}{\int_0^\infty w(y, \theta) dF_1(y)} \right]^{r_{2j}} \\ = & \prod_{j=1}^{h} p_j^{r_{1j}+r_{2j}} \left[\int_0^\infty w(y, \theta) dF_1(y) \right]^{-n_2} \\ & \cdot \prod_{j=1}^{h} w^{r_{2j}}(t_j, \theta) \\ = & \prod_{j=1}^{h} p_j^{r_j} \left[\sum_{j=1}^{h} w(t_j, \theta) p_j \right]^{-n_2} \prod_{j=1}^{h} w^{r_{2j}}(t_j, \theta). \end{aligned}$$

Thus, the log-likelihood equations are

$$\frac{\partial l(\theta, \boldsymbol{p})}{\partial p_j} = \frac{r_j}{p_j} - \frac{n_2 w(t_j, \theta)}{\sum_{\ell=1}^{h} w(t_\ell, \theta) p_\ell} = 0$$

for $j = 1, \ldots, h$. Let

$$V(\theta, \boldsymbol{p}) = \sum_{\ell=1}^{h} w(t_\ell, \theta) p_\ell.$$

It follows that

$$\begin{aligned} p_j &= \frac{r_j}{n_2 w(t_j,\theta)} \sum_{\ell=1}^{h} w(t_\ell,\theta) p_\ell \\ &= \frac{r_j}{n_2 w(t_j,\theta) V^{-1}(\theta,\boldsymbol{p})}. \end{aligned}$$

Since $\sum_{j=1}^{h} p_j = 1$, we have

$$V^{-1}(\theta,\boldsymbol{p}) = \sum_{j=1}^{h} \frac{r_j}{n_2 w(t_j,\theta)}.$$

That is, for $j = 1,\ldots,h$,

$$p_j = \left[\frac{r_j}{w(t_j,\theta)}\right] / \left[\sum_{\ell=1}^{h} \frac{r_\ell}{w(t_\ell,\theta)}\right]. \qquad (1)$$

Substituting (1) into $L(\theta,\boldsymbol{p})$ to get the profile partial likelihood

$$\begin{aligned} L_{pr} &= L(\theta, p(\theta)) \\ &= \prod_{j=1}^{h} \left[\frac{\frac{r_j}{w(t_j,\theta)}}{\sum_{\ell=1}^{h} \frac{r_\ell}{w(t_\ell,\theta)}}\right]^{r_j} \left[\frac{n_1+n_2}{\sum_{\ell=1}^{h} \frac{r_\ell}{w(t_\ell,\theta)}}\right]^{-n_2} \\ &\qquad \cdot \prod_{j=1}^{h} w^{r_{2j}}(t_j,\theta) \\ &= (n_1+n_2)^{-n_2} \left\{\prod_{j=1}^{h} \left[\frac{r_j}{w(t_j,\theta)}\right]^{r_j}\right\} \\ &\qquad \cdot \left[\sum_{\ell=1}^{h} \frac{r_\ell}{w(t_\ell,\theta)}\right]^{-n_1} \prod_{j=1}^{h} w^{r_{2j}}(t_j,\theta) \\ &= (n_1+n_2)^{-n_2} \left[\prod_{j=1}^{h} \frac{r_j^{r_j}}{w^{r_{1j}}(t_j,\theta)}\right] \\ &\qquad \cdot \left[\sum_{\ell=1}^{h} \frac{r_\ell}{w(t_\ell,\theta)}\right]^{-n_1} \end{aligned}$$

Then

$$\begin{aligned} & l_{pr}(\theta) = \log L_{pr}(\theta) \\ = & -\sum_{j=1}^{h} r_{1j} \log w(t_j,\theta) \\ & -n_1 \log\left[\sum_{\ell=1}^{h} \frac{r_\ell}{w(t_\ell,\theta)}\right] + \text{ Constant.} \end{aligned}$$

A special case with $w(t,\theta) = \exp(\theta t)$ gives

$$\begin{aligned} l_{pr}(\theta) &= \text{Constant} - \sum_{j=1}^{h} r_{1j}\theta t_j \\ &\quad -n_1 \log\left[\sum_{j=1}^{h} r_j \exp(-\theta t_j)\right]. \end{aligned}$$

If there is no tie among the observations, then

$$\begin{aligned} l_{pr}(\theta) &= \text{Constant} - \sum_{j=1}^{n_1} \theta X_j \\ &\quad -n_1 \log\left[\sum_{j=1}^{n_1+n_2} \exp(-\theta t_j)\right] \\ &= \text{Constant} - \sum_{j=1}^{n_1} \theta X_j \\ &\quad -n_1 \log\left[\sum_{j=1}^{n_1} \exp(-\theta X_j) \right. \\ &\qquad \left. + \sum_{j=1}^{n_2} \exp(-\theta Y_j)\right]. \end{aligned}$$

Now we move to the case of two-sample censored data. For $i = 1,\ldots,n_1$ and $j = 1,\ldots,n_2$, assume that the censored times C_i and D_j are distributed as G_1 and G_2. The observed censored data is (X_i,Δ_i) and (Y_j,Γ_j) , where $X_i = \min(X_i^0, C_i)$, $Y_j = \min(Y_j^0, D_j)$, $\Delta_i = I(X_i^0 < C_i)$, $\Gamma_j = I(Y_j^0 < D_j)$. It is obvious that the log-likelihood (incomplete data, p^{old},θ^{old})is linear in r_{1j} and r_{2j}. So we only need to replace r_{1j} and r_{2j} by $E(r_{1j}|\text{incomplete data}, p^{\text{old}}, \theta^{\text{old}})$ and $E(r_{2j}|\text{incomplete data}, p^{\text{old}}, \theta^{\text{old}})$. Then we can repeat the maximization as complete data. Recall that

$$r_{1j} = \sum_{i=1}^{n_1} I(X_i^0 = t_j),$$

and

$$r_{2j} = \sum_{i=1}^{n_2} I(Y_i^0 = t_j),$$

then

$$
\begin{aligned}
& E(r_{1j}|\text{incomplete data}, p^{\text{old}}, \theta^{\text{old}}) \\
= & \sum_{i=1}^{n_1} E\left[I(X_i^0 = t_j)|(X_i, \Delta_i)_{i=1}^{n_1},\right. \\
& \left.(Y_i, \Gamma_i)_{i=1}^{n_2}, p^{\text{old}}, \theta^{\text{old}}\right] \\
= & \sum_{i=1}^{n_1} E\left[I(X_i^0 = t_j, \Delta_i = 1)|\right. \\
& \left.(X_i, \Delta_i) = (x_i, \delta_i), p^{\text{old}}, \theta^{\text{old}}\right] \\
& + \sum_{i=1}^{n_1} E\left[I(X_i^0 = t_j, \Delta_i = 0)|\right. \\
& \left.(X_i, \Delta_i) = (x_i, \delta_i), p^{\text{old}}, \theta^{\text{old}}\right] \\
= & \sum_{i=1}^{n_1} (I_{i1} + I_{i2}).
\end{aligned}
$$

Now $I_{i1} = I(x_i = t_j, \delta_i = 1) = \#$ of uncensored deaths at t_j from sample $1 \equiv \xi_{1j}$. Also,

$$
\begin{aligned}
I_{i2} &= I(\delta_i = 0) I(x_i \le t_j) \\
&\quad \frac{P(X_i^0 = t_j, C_i = x_i | p^{\text{old}}, \theta^{\text{old}})}{\sum_{\ell: t_\ell \ge x_i} P(X_i^0 = t_\ell, C_i = x_i | p^{\text{old}}, \theta^{\text{old}})} \\
&= I(\delta_i = 0) I(x_i \le t_j) \\
&\quad \frac{dF_1^{\text{old}}(t_j) dG_1(x_i)}{\sum_{\ell: t_\ell \ge x_i} dF_1^{\text{old}}(t_\ell) dG_1(x_i)} \\
&= I(x_i \le t_j, \delta_i = 0) \frac{p_j^{\text{old}}}{\sum_{\ell: t_\ell \ge x_i} p_\ell^{\text{old}}}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& \sum_{i=1}^{n_1} I_{i2} \\
= & \sum_{i=1}^{n_1} I(x_i \le t_j, \delta_i = 0) \frac{p_j^{\text{old}}}{\sum_{\ell: t_\ell \ge x_i} p_\ell^{\text{old}}} \\
= & p_j^{\text{old}} \sum_{k=1}^{j} \sum_{i=1}^{n_1} \frac{I(x_k = t_j, \delta_i = 0)}{\sum_{\ell: t_\ell \ge x_i} p_\ell^{\text{old}}} \\
= & p_j^{\text{old}} \sum_{k=1}^{j} \frac{\sum_{i=1}^{n_1} I(x_k = t_j, \delta_i = 0)}{\sum_{\ell: t_\ell \ge x_i} p_\ell^{\text{old}}} \\
= & p_j^{\text{old}} \sum_{k=1}^{j} \frac{\#\text{ of censored } x_i\text{'s at } t_k \equiv \eta_{1k}}{\sum_{\ell: t_\ell \ge x_i} p_\ell^{\text{old}}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& E(r_{1j}|\text{incomplete data}, p^{\text{old}}, \theta^{\text{old}}) \\
= & \xi_{1j} + p_j^{\text{old}} \sum_{k=1}^{j} \frac{\eta_{1k}}{\sum_{\ell=k}^{h} p_\ell^{\text{old}}}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& E(r_{2j}|\text{incomplete data}, p^{\text{old}}, \theta^{\text{old}}) \\
= & \xi_{2j} + p_j^{\text{old}} \sum_{k=1}^{j} \frac{\eta_{2k}}{\sum_{\ell=k}^{h} p_\ell^{\text{old}}},
\end{aligned}
$$

where ξ_{2j} and η_{2j} are defined similarly as ξ_{1j} and η_{1j}. Thus, in the $s+1$th step of the algorithm,

$$
\begin{aligned}
& p_j^{(s+1)}(\theta) \\
= & \frac{E(r_j|data, p^{(s)}, \theta^{(s)})/w(t_j, \theta^{(s)})}{\sum_{i=1}^{h} E(r_i|data, p^{(s)}, \theta^{(s)})/w(t_i, \theta^{(s)})} \\
= & \frac{\left[\xi_j + p_j^{(s)}(\theta^{(s)}) \sum_{k=1}^{j} \frac{\eta_k}{\sum_{\ell=k}^{h} p_\ell^{(s)}}\right] / w(t_j, \theta^{(s)})}{\sum_{i=1}^{h} \left[\xi_i + p_i^{(s)}(\theta^{(s)}) \sum_{k=1}^{i} \frac{\eta_k}{\sum_{\ell=k}^{h} p_\ell^{(s)}}\right] / w(t_i, \theta^{(s)})}.
\end{aligned}
$$

for $j = 1, \ldots, h$. Then $\theta^{(s+1)}$ is the value that maximizes the log-likelihood function $l_C = \log L_C$ which is slightly different from the log-likelihood function of complete data, where

$$
\begin{aligned}
L_C &= \prod_{i=1}^{n_1} [dF_1(x_i)]^{\Delta_i} [1 - F_1(x_i)]^{1-\Delta_i} \\
&\quad \cdot \prod_{j=1}^{n_2} [dF_2(y_j)]^{\Gamma_j} [1 - F_2(y_j)]^{1-\Gamma_j} \\
&= \prod_{j=1}^{h} [dF_1(t_j)]^{\xi_{1j}} [1 - F_1(t_j)]^{\eta_{1j}} \\
&\quad \prod_{j=1}^{h} [dF_2(t_j)]^{\xi_{2j}} [1 - F_2(t_j)]^{\eta_{2j}} \\
&= \prod_{j=1}^{h} p_j^{\xi_{1j}} \left(\sum_{k=j}^{h} p_k\right)^{\eta_{1j}} \\
&\quad \cdot \prod_{j=1}^{h} \left[\frac{w(t_j, \theta) p_j}{\sum_{i=1}^{h} w(t_i, \theta) p_i}\right]^{\xi_{2j}}
\end{aligned}
$$

$$\cdot \left[\frac{\sum_{k=j}^{h} w(t_k,\theta)p_k}{\sum_{i=1}^{h} w(t_i,\theta)p_i}\right]^{\eta_{2j}}$$

$$= \prod_{j=1}^{h}\left\{p_j^{\xi_{1j}+\xi_{2j}}\left(\sum_{k=j}^{h} p_k\right)^{\eta_{1j}} w^{\xi_{2j}}(t_j,\theta)\right.$$
$$\cdot \left[\sum_{k=j}^{h} w(t_k,\theta)p_k\right]^{\eta_{2j}}$$
$$\left.\cdot \left[\sum_{i=1}^{h} w(t_i,\theta)p_i\right]^{-(\xi_{2j}+\eta_{2j})}\right\}$$

and

$$l_C = \sum_{j=1}^{h}\left\{(\xi_{1j}+\xi_{2j})\log p_j + \eta_{1j}\log\left(\sum_{k=j}^{h} p_k\right)\right.$$
$$+\xi_{2j}\log w(t_j,\theta) + \eta_{2j}\log\left[\sum_{k=j}^{h} w(t_k,\theta)p_k\right]$$
$$\left.-(\xi_{2j}+\eta_{2j})\log\left[\sum_{i=1}^{h} w(t_i,\theta)p_i\right]\right\}.$$

The algorithm terminates the searching process until we reach a convergence. The speed of this searching process is quite slow when the censoring rates are as low as 20% on both samples. We have simulated the 20% and 20%, 20% and 50%, 50% and 20%, and 50% and 50% four combinations of censored rates for $n_1 = n_2 = 50$ and $n_1 = n_2 = 100$. Our estimates are performing pretty good and we shall try to compare our method with others in the literature in the future.

2 References

Gilbert, P.B., Lele, S.R., Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* **86**, 27–42.

Sun, J. and Woodroofe, M. (1997). Semiparametric estimates under biased sampling. *Statistica Sinica* **7**, 545–575.

Andersen, P.K. and Vaeth, M. (1989). Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* **45**, 523–535.

Li, G. and Qin, J. (1999). Semiparametric analysis of two-sample truncated data. Submitted for publication.

Chen, W.P. (1999). To compare the survival rates in invasive breast cancer among ethnically Chinese women born in east Asian and the unites states. Master Report. Department of Biostatistics, UCLA.